DOCUMENT RESUME

ED 234 074 TM 830 558

AUTHOR Harnisch, Delwyn L.
TITLE An Investigation of the Interaction between Item Content and Group Membership on the Tests of General Educational Development.
PUB DATE Apr 83
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Aptitude Treatment Interaction; *Equivalency Tests; High School Equivalency Programs; *Item Analysis; *Latent Trait Theory; National Norms; Research Methodology; Sampling; Secondary Education; *Student Characteristics; Test Items; Writing Evaluation
IDENTIFIERS *General Educational Development Tests; *Three Parameter Model

ABSTRACT

The purpose of this investigation was to examine the item characteristics of the General Educational Development Tests (GED) for specific subgroups of students. The method employed was the approach suggested by Linn and Harnisch in which the three-parameter logistic model is used for detecting interactions between item content and group membership. Data used in this investigation came from four separate forms of the GED writing test administered to national representative samples during the spring term of the 1979-1980 testing cycle. Results of the study are summarized for the five content categories (spelling, capitalization/punctuation, usage, sentence completion, and logic and organization) on the writing test for female students for each of the four forms. The discussion of the results suggest that the interactions between item content and group membership may be conceptualized as indications of multidimensionality. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED234074

An Investigation of the Interaction Between Item Content and Group Membership on the Tests of General Educational Development

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Delwyn L. Harnisch

University of Illinois, Urbana-Champaign

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D. Harnisch

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 830 558

Paper presented at the annual meeting of the American Educational Research Association, Montreal, April, 1983.

An Investigation of the Interaction Between Item Content and Group Membership on the Tests of General Educational Development

The purpose of this investigation was to examine the item characteristics for specific subgroups of students. There are a wide variety of student characteristics that could interact with item characteristics to affect overall performance on a test. Four potentially important characteristics include ethnic group membership, sex, socioeconomic status and differences in motivation or test taking strategies.

Background

It has been widely documented repeatedly that "...children of the poor tend to score lower... on standardized ability and achievement tests than do children of the affluent..." (Stanley, 1971, p. 640). Numerous studies have shown that members of some minority groups tend to score lower on standardized ability and achievement tests than do members of the white majority. Groups of children who differ in terms of motivation, test taking anxiety, or in their success and failure attributions also differ in their average performance on achievement tests. Performance differences are also to be expected for groups that have had differential exposure to the content.

Group differences in test performance taken alone tell us nothing about the reasons for those differences. The differences alone cannot even distinguish between the possibility that the results are reflecting real and important differences in the knowledge and understanding of the subject matter covered by the test and the possibility that the results reflect inadequacies in

the measuring instrument. That is to say that the test might be biased against one of the groups.

Test results must be interpretable to have any utility (Linn & Harnisch, 1981). When test scores on a writing test are interpreted as an indicator of competence in writing, a host of alternative explanations must be ruled out either explicitly or as is more often the case, implicitly. The examination of alternative explanations of test performance often involves the investigation of possible interactions. For example, an interaction between speediness of a test and student anxiety (Hill, 1980) may indicate that a test given under its standardized timing conditions yields biased estimates of the competence of high anxious children. The focus of this investigation is on the interaction between characteristics of examinees and characteristics of items.

It is often argued that certain types of items are biased against some groups (for example, sex, race/ethnicity, age, culture, etc.). For example, in a review by Phyliss Rosser of the Differential Aptitude Test published by Harcourt Brace Jovanovich she notes that "the mechanical reasoning test contains content that is highly male-oriented; pictures of men looking in mirrors, riding horses, throwing a ball, using a hose and standing on the moon. Meanwhile, the only woman on the test is sitting in a wheelchair"(Rosser, 1981, p. 4). She further notes that on the language usage section of the test that male nouns and pronouns outnumber females three to one.

It is less obvious for many types of achievement tests to identify the types of item characteristics that might interact

with student characteristics. To identify the student characteristics which might interact with a particular content area of problems is less apparent than the identification of items dealing with cultural contributions. Thus a more exploratory approach is needed to examine the interaction between the characteristics of items and characteristics of students.

Over 30 years ago researchers noted the need to identify characteristics which might interact with student characteristics. Previous efforts have fallen short in identifying general characteristics of test items that result in underestimating the ability level of specific subgroups of students. The earlier results (Angoff & Ford, 1973; Cleary & Hilton, 1968; Green, 1973; Jensen, 1974) were not very informative since they were based entirely on sample dependent statistics.

Item response theory (IRT) provides a more theoretically sound approach to the problems of identifying items that function differently for different groups. One of the distinct advantages of the use of IRT models is that they provide invariant item parameters. Recent IRT studies examining item bias (Ironson & Subkoviak, 1979; Rudner & Convey, 1978; Rudner, Getson & Knight, 1980; Shepard, Camilli & Averill, 1980) have compared the use of latent trait models to identify biased items to a variety of techniques that had previously been used. Some agreement among the techniques was found but it was clear from the results that the more traditional techniques are not substitutes for an approach based on latent theory.

One of the advantages of the latent trait model approach is

the possibility of providing invariant item parameters. Wright (1977) often refers to this as "person-free item calibration." That is to say that if a given latent trait model is used then the estimated item parameters obtained independently for the two groups should be the same, except for sampling error, once they have been put on the same scale.

The preference for the three-parameter model is natural when one has sufficiently large samples in each group. For each group estimates would be obtained separately. By using a linear equating of the difficulty parameter estimates the estimates would be placed on the same scale. Following the outline as suggested by Lord (1977, 1980) comparisons of the item characteristic curves would be made. However, since this approach requires large samples (following Lord's guidelines for the use of LOGIST a sample of 1000 would be needed in the smallest of the groups being compared) an alternative approach suggested by Linn and Harnisch (1981) was used.

Method

The Linn and Harnisch (1981) approach starts by obtaining the estimates of the item discriminating power, a, the item difficulty, b, and the lower asymptote,c, of the three-parameter logistic model based upon all available cases in the sample. The results from LOGIST also provided an estimate of each person's location along the latent trait, θ. Based on these estimates the estimated probability that person j would answer item i correctly were obtained in the usual way. That is,

$$P_{ij} = C_i + \frac{1 - C_i}{1 + \exp\left[-1.7a_i(\theta_j - b_i)\right]} \quad ,$$

where a , b , c and are all estimates. These estimates based on the model when averaged over members of a subgroup can be compared to the observed proportion correct for that subgroup. If person j is a member of group g, then the proportion of people in group g expected to get item i correct according to the model is

$$P_{ig} = \frac{1}{n_g} \sum_{j \in g} P_{ij} \quad ,$$

where n is the number of persons in group g. The observed proportion correct on item i for group g, O , is simply the number of people in group g who answer item i correctly divided by n . The difference,

$$D_i = O_i - P_i \quad ,$$

is an index of the degree to which members of that subgroup perform better or worse than expected on that item. The difference and a standardized difference between observed and expected performance was used for members of various subgroups to identify items that are unusually easy or difficult for members of those subgroups. The standardized difference score for

subgroup g is defined as

$$Z_{ig} = \frac{1}{n_g} \sum_{j \in g} \left[\frac{U_{ij} - P_{ij}}{\sqrt{P_{ij}\,(1 - P_{ij})}} \right] ,$$

where $U_{ij} = 1$, when person j answers item i correctly and $U_{ij} = 0$ otherwise. Thus, the standardized difference for the complete target group is

$$Z_{i.} = \sum_g n_g Z_{ig} \Big/ \sum_g n_g \quad .$$

Data Source

The data used in this investigation came from four separate forms of the General Educational Development (GED) mathematics test and four separate forms of the GED writing test administered to national representative samples during the spring term of the 1979-1980 testing cycle. Each form of both the mathematics and writing tests had approximately 1100 students responding to each of the 50 cognitive items of mathematics and 80 writing cognitive items. The forms for both tests were selected based on the representativeness of the respondents to these respective forms. Specifically, the sample sizes for the four forms (34-37) of the mathematics test were 1194, 1123, 880 and 1059 respectively while the sample sizes for the four forms (34-37) of the writing test were 1458, 1446, 1124 and 1158 respectively.

Results and Discussion

In the time remaining, it is impossible to report on the

results of both tests for all forms included in the GED test. Nor is it possible to give results for the eight subgroups that were investigated and are continuing to investigate. The eight subgroups represent two sex groups, two race/ethnic groups and four age groups. Instead I will focus on the writing test with one form for one subgroup, female students. Thus, the results may be taken as illustrative of the type that may be obtained on other forms of the writing test as well as on other tests and with other groups.

The writing test contains 80 items. Each form of the writing test includes five types of items. The item types and number of items noted in parenthesis include spelling (10), capitalization/punctuation (10), usage (24), sentence completion (24) and logic and organization (12). The apparent differences in the content categories was not found to deviate substantially from the unidimensionality of the items that is assumed by the latent trait models. Factor analysis of the subtest scores revealed one major factor accounting for the major proportion of the variance for each of the tests. A scree test was performed to acknowledge the fact that unidimensionality was approximated by each of the collection of items on each form. Of greater interest for this investigation with particular subgroups is the extent to which items and groups of items show systematic differences in the direction and magnitude of the observed versus expected proportion correct as indicated by the difference or standardized difference index.

The estimates of the parameters of the three parameter logistic model were obtained using LOGIST (Wood, Wingersky &

Lord, 1976) for the sample of all students taking Form 37 of the writing test with usable data (N = 1092). The results were then used to obtain estimates of the expected chance that each student would have of getting each item right, P_{ij}. These estimates were compared to the observed results separately for male and for female students. The comparisons for the 591 female students in the sample were made as a function of estimated values of θ. First the female students were divided into quintiles on the basis of their estimated θ's. Within each quintile, P_{ig}, O_{ig} and D_{ig} was then computed for each item. This allowed for the possibility that students at one level of θ may perform systematically better than expected on an item while those at another level perform systematically worse than expected.

Average standardized difference scores, Z_{ig}, were also computed within each quintile as well as across all female students. The latter was used as an overall index for each item.

The expected and observed proportion correct for the five quintiles of two items is shown in Figures 1 and 2. The points, P_{ig} and O_{ig} are plotted above the mean theta value within each quintile. The solid lines connect the observed proportions, P_{ig}, and the dashed lines connect the expected proportions. The item shown in Figure 1 has the largest positive standardized difference score averaged all quintiles. The item depicted in Figure 2 has the largest negative standardized difference score averaged over all quintiles.

Items like the one shown in Figure 1 may be considered biased in favor of female students because their actual performance is systematically better than expected relative to

other items in the test. While on the other hand items like the one shown in Figure 2 may be considered biased against female students relative to other items on the test.

The results in Figure 1 and 2 represent the extremes. For these items the differences are rather substantial. The performance of female students would appear better on a test that had more items of the type shown in Figure 1 and/or fewer items of the type shown in Figure 2. However, these results by themselves do not indicate what leads to these differences. To examine some possible reasons for these differences we categorized items into content categories and then compared the items in different categories in terms of their average standardized difference scores, Z_i.

Of the five items with the largest positive standardized differences, three of them involved questions about spelling while the other two were representatives from the category of sentence completion and from the category of capitalization/punctuation. In contrast, none of the five items on which the actual performance of female students was less than expected by the greatest amount (as measured by the largest Z_i's) involved spelling. Two were language usage, one was from the logic and organization category while the other two were from the sentence completion category. The item depicted in Figure 1 is from the spelling category while the item depicted in Figure 2 is an item from the sentence completion category. These results are hardly conclusive but are rather suggestive.

After categorizing the items by content, the patterns of standardized difference scores were quite distinctive for some of

the categories. For example, of the 10 spelling items, nine of them had positive Z_i's and the tenth one was -.012. Similarly, 17 of the 24 usage problems had negative Z's. The results for the remaining three categories were less consistent and have small positive or negative mean Z values near zero.

The mean Z and range of Z's within each category of items are listed in Table 1. The Z values for similar content classifications across the remaining three forms of the writing test are also given in Table 1. Spelling problems across all four forms reveal the same central tendency for the Z values in the positive direction. Logic and organization problems stand out as problems in the other direction for all four forms which is to say that females tend not to perform as well as expected from the model on these problems.

Differences such as those noted above are by themselves interesting and might be labelled item bias. To note that the spelling items appear to be "biased" in favor of female students while logic and organization items are "biased" against them is not very helpful. The model implies that the items are at fault while it is at least as plausible to consider that the model might be at fault.

Since the GED writing test contains the same number of items of each category across forms it seems clear that the gender differences will appear to be nearly the same across forms. However, with future revisions of the GED any modification of the number of spelling or logic and organization items would be expected to alter the magnitude of gender differences on the global score in predictable ways.

Finally, a review of these items by other approaches such as judgmental by outside specialists would help to validate the interaction of student characteristics with item content. Furthermore, in situations where the judgmental review does not match the statistical review it may well be worth the expense of both approaches to note that the two procedures are identifying two or more different item bias concerns.

References

Angoff, W. H., & Ford, S. F. Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-106.

Cleary, T. A., & Hilton, T. L. An investigation of item bias. Educational and Psychological Measurement, 1968, 28, 61-75.

Green, D. R. Racial and ethnic bias in achievement tests and what to do about it. Monterey, California: CTB/McGraw-Hill, 1973.

Hill, K. T. Motivation, evaluation and educational testing policy. In L. J. Fyans, Jr. (Ed.), Achievement motivation: Recent trends in theory and research. New York: Plenum Press, 1980.

Ironson, G. H. & Subkoviak, M. A comparison of several methods of assessing item bias. Journal of Educational Measurement, 1979, 16, 209-225.

Jensen, A. R. How biased are culture-loaded tests? Genetic Psychology Monographs, 1974, 40, 185-244.

Linn, R. L. & Harnisch, D. L. Interactions between item content and group membership on achievement test items, Journal of Educational Measurement, 1981, 18, 109-118.

Lord, f. M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates, 1980, Chapter 14.

Lord, F. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1974, 11, 235-245.

Rosser, P. Sex bias in testing. The Testing Digest, 1981, 3, 1-6.

Rudner, L. M. & Convey, J. J. An evaluation of select approaches for biased item identification. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March, 1978.

Rudner, L. M., Getson, P. R. & Knight, D. L. A Monte Carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 1980, 17, 1-10.

Shepard, L., Camilli G. & Averill, M. Comparison of six procedures for detecting test item bias using both internal and external ability criteria. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, April, 1980.

Stanley, J. C. Predicting college success of the educationally disadvantaged. Science, 1971, 171, 640-647.

Wood, R. L., Wingersky, M. S. & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. (ETS RM 76-6). Princeton, NJ: Educational Testing Service, 1976.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Acknowledgements

The author wishes to thank the staff of the GED testing service of the American Council on Education for the preparation of the data tape used in this investigation. Special thanks are particularly extended to Douglas R. Whitney, Wayne M. Patience and Andrew G. Malizio for their support and assistance in the development of the data files. The author is also grateful to Stephen Dunbar for his assistance with the development of necessary computer programs and for carrying out various computer analyses.

Table 1

The Mean and Range of Standardized Differences

Between Observed and Expected

Proportion Correct by Content Categories

Form	Content Category	Number of Items	Mean Z	Range of Z's
34	Spelling	10	.05	-.03 to .10
	Capitalization/Punctuation	10	.00	-.06 to .09
	Usage	24	-.02	-.08 to .04
	Sentence Completion	24	-.01	-.05 to .05
	Logic and Organization	12	-.04	-.07 to .00
35	Spelling	10	.04	-.02 to .08
	Capitalization/Punctuation	10	.00	-.07 to .05
	Usage	24	-.01	-.08 to .04
	Sentence Completion	24	.00	-.10 to .08
	Logic and Organization	12	-.06	-.11 to .00
36	Spelling	10	.05	-.01 to .11
	Capitalization/Punctuation	9*	.01	-.05 to .08
	Usage	24	-.01	-.09 to .09
	Sentence Completion	24	-.01	-.08 to .10
	Logic and Organization	12	-.06	-.10 to .02
37	Spelling	10	.06	-.01 to .15
	Capitalization/Punctuation	10	.03	-.05 to .10
	Usage	24	-.03	-.11 to .05
	Sentence Completion	24	-.01	-.14 to .09
	Logic and Organization	12	-.01	-.10 to .07

* One item was deleted from the analysis because it was not fitting the latent trait model.

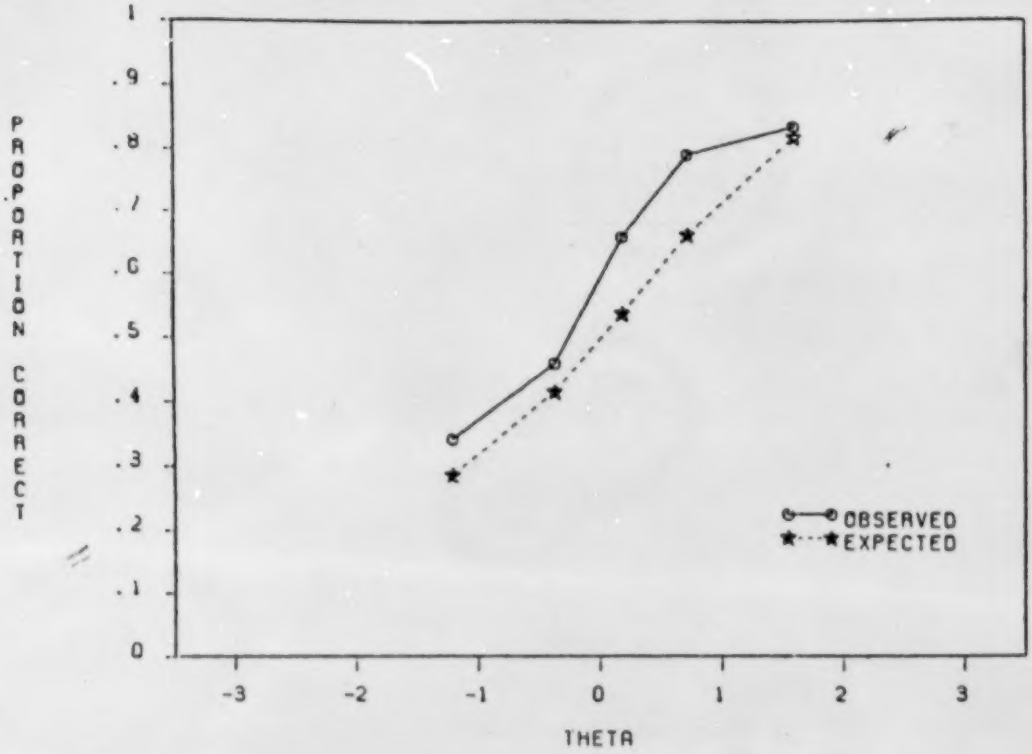

Figure 1. Observed and expected proportion correct as a function of theta for an item with a large positive difference.

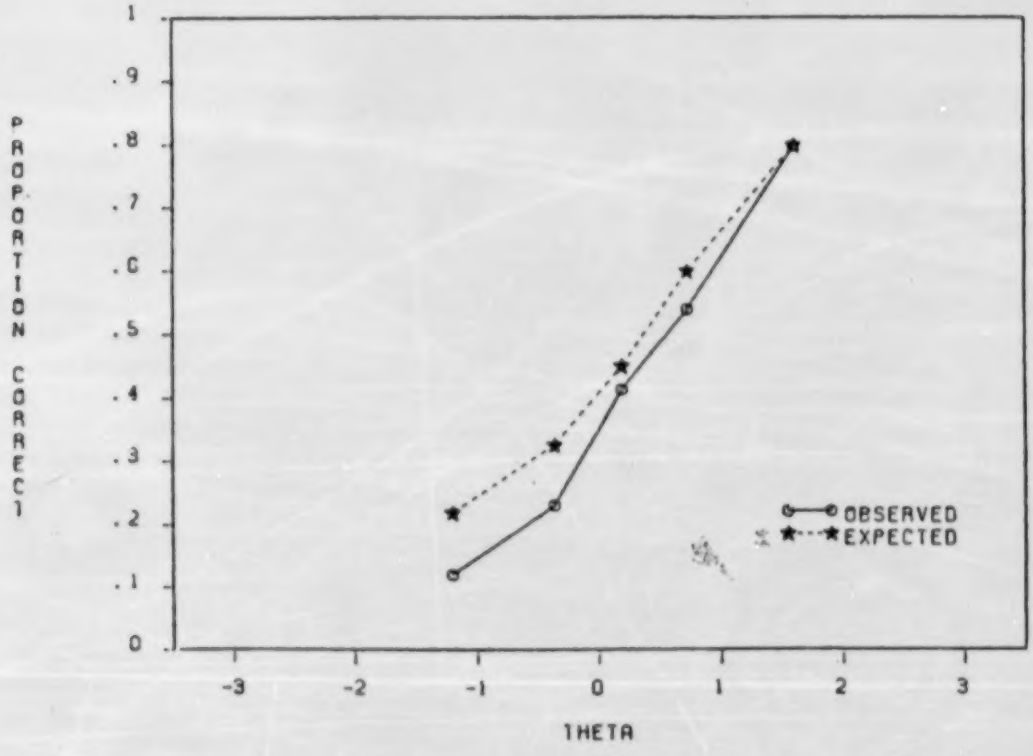

Figure 2. Observed and expected proportion correct as a function of theta for an item with a large negative difference.